BMC Bioinformatics

RESEARCH

Open Access

Improved quality metrics for association and reproducibility in chromatin accessibility data using mutual information

Cullen Roth[1*], Vrinda Venu[2], Vanessa Job[3], Nicholas Lubbers[4], Karissa Y. Sanbonmatsu[5], Christina R. Steadman[2] and Shawn R. Starkenburg[1]

*Correspondence:
croth@lanl.gov

[1] Los Alamos National Laboratory, Genomics and Bioanalytics, Los Alamos, NM, USA
[2] Los Alamos National Laboratory, Climate, Ecosystems, and Environmental Science, Los Alamos, NM, USA
[3] Los Alamos National Laboratory, High Performance Computing and Design, Los Alamos, NM, USA
[4] Los Alamos National Laboratory, Information Sciences, Los Alamos, NM, USA
[5] Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, NM, USA

Abstract

Background: Correlation metrics are widely utilized in genomics analysis and often implemented with little regard to assumptions of normality, homoscedasticity, and independence of values. This is especially true when comparing values between replicated sequencing experiments that probe chromatin accessibility, such as assays for transposase-accessible chromatin via sequencing (ATAC-seq). Such data can possess several regions across the human genome with little to no sequencing depth and are thus non-normal with a large portion of zero values. Despite distributed use in the epigenomics field, few studies have evaluated and benchmarked how correlation and association statistics behave across ATAC-seq experiments with known differences or the effects of removing specific outliers from the data. Here, we developed a computational simulation of ATAC-seq data to elucidate the behavior of correlation statistics and to compare their accuracy under set conditions of reproducibility.

Results: Using these simulations, we monitored the behavior of several correlation statistics, including the Pearson's R and Spearman's ρ coefficients as well as Kendall's τ and Top–Down correlation. We also test the behavior of association measures, including the coefficient of determination R^2, Kendall's W, and normalized mutual information. Our experiments reveal an insensitivity of most statistics, including Spearman's ρ, Kendall's τ, and Kendall's W, to increasing differences between simulated ATAC-seq replicates. The removal of co-zeros (regions lacking mapped sequenced reads) between simulated experiments greatly improves the estimates of correlation and association. After removing co-zeros, the R^2 coefficient and normalized mutual information display the best performance, having a closer one-to-one relationship with the known portion of shared, enhanced loci between simulated replicates. When comparing values between experimental ATAC-seq data using a random forest model, mutual information best predicts ATAC-seq replicate relationships.

Conclusions: Collectively, this study demonstrates how measures of correlation and association can behave in epigenomics experiments. We provide improved strategies for quantifying relationships in these increasingly prevalent and important chromatin accessibility assays.

Keywords: ATAC-seq, Correlation and association, Normalized mutual information, Random forest

Background

Epigenetic modifications play an important role in regulating multiple cellular processes ranging from DNA replication to gene expression. These covalent additions to DNA and histone proteins do not alter the underlying DNA sequence, but rather, help modulate chromatin structure resulting in distinctive phenotypes. Genome-wide epigenetic modifications can be determined using several techniques: the gold-standard is chromatin immunoprecipitation followed by sequencing (ChIP-seq) [1–3]. Chromatin accessibility, or the analysis of the regions that are available for DNA:protein interactions potentially resulting in gene expression, is measured using an enzyme-driven assay called transposase-accessible chromatin via sequencing (ATAC-seq) [4]. These two methods have distinct advantages in probing the state of the epigenome, and both approaches generate paired-end sequencing libraries. These data are mapped to the genome to determine the loci that are occupied with a particular epigenetic modification or the loci that are localized within an open, accessible region. Epigenetic modifications and chromatin accessibility are visualized as peaks resulting from the aggregation of sequencing reads [5]. As such, many software platforms used for analysis of ChIP-seq and ATAC-seq data sets use 'peak calling' to determine locations of epigenetic modifications or accessible chromatin regions [6–9].

To ensure significance and consistency of identified peaks, best practices have been defined for quantifying reproducibility across experimental replicates [8, 10]. These include several quality control metrics and workflows that standardize analysis and enable comparison among different experiments [10]. These standards apply to the total number of sequenced reads, total number of identified significant peaks, and concentration of sequenced reads within said peaks. For example, pseudo-replication was developed for ChIP-seq analysis to assess the amount of variation between biological replicates [8]. In this protocol, synthetic replicates are created from true, experimentally derived data: to do this, aligned reads are merged from two true replicates and randomly reassigned into new alignments to create two synthetic replicates. This permutation practice homogenizes (and splits) signals present within the true, observed replicates, generating the null hypothesis of near perfect correlation between pseudo-replicates. Peak calling is then also conducted on pseudo-replicates, and the read counts of peaks conserved between the two pseudo-replicates are compared to the observed peaks in the true replicates. Landt et al. proposed that experiments, whose number of observed peak counts (among true replicates) divided by the total number of pseudo peaks (between pseudo-replicates), which nears a value of one, are broadly reproducible [8]. The ENCODE project has since extended this practice to ATAC-seq experiments [11, 12].

To better understand experimental reproducibility, many studies also conduct correlation analysis on binned signals between ATAC-seq replicates [13–15]. In such analyses, for each replicate, the genome is binned into smaller, contiguous regions, for example using windows of ten kilobase pairs [13]. The number of mapped sequenced fragments (defined by a pair of mapped reads) that overlap these bins are counted and standardized to fragments per kilobase pair per million reads (Fpkm) [16]. These Fpkm counts are then compared between replicates using correlation and association statistics such as Pearson's R or the coefficient of determination (R^2), respectively. Values from these statistics trending toward a value of one generally indicate a reproducible experiment [17].

Correlation analysis is a useful tool, not singularly purposed for the analysis of reproducibility in ATAC-seq experiments. Such analysis can be found within studies of chromosome accessibility in cancer, ageing of human stem cells, cellular diversity, or new ATAC-seq protocols [18–23]. Furthermore, correlation analyses are ubiquitous, found in the fields of genetics, RNA-seq experiments, and in studies of 3D chromatin architecture [16, 24–30]. Given their popularity and use in genomic and epigenetic studies, software suites—for example deeptools and HiCExplorer—have developed methods and tools for calculating correlation metrics between replicates and experiments [13, 31–34].

The natural properties of data from genomic and epigenomic experiments make the application of commonly used correlation and association statistics, for example Pearson's R and R^2, potentially problematic as none of these data (ATAC-, ChIP-, or Hi-C seq) are normally distributed [35]. Both ATAC- and ChIP-seq experiments are defined by numerous, loci-specific peaks of signal generated by the accumulation of sequencing reads [3, 4]. Mapped sequenced fragments may overlap contiguous genomic bins used in analysis, producing non-independent data points [24]. Conversely, regions lacking assayed modifications or with inaccessible chromatin will have little to zero signal for ChIP-seq or ATAC-seq data, respectively. Furthermore, during correlation analysis, several genomic bins may overlap an inaccessible chromatin region that is reproducible, appearing in both the ATAC-seq replicates (or experiments) being compared. As such, each of these bins will acquire zero Fpkm and within the bi-variate distribution formed between the replicates. These data points, which appear as zero Fpkm in both replicates, are referred to here as co-zeros. Some analysis programs, like deeptools, HiCExplorer, and HiCcompare, offer options to remove co-zeros prior to analysis [29, 31, 34]. However, there is no published guidance on this practice, and while the co-zero values are a feature common across genomic and epigenomic data sets [36], the effect of removing such features on correlation statistics has not been explored. Despite the known features of genomic and epigenomic data, and the underlying assumptions of statistical tests, there have been few studies that explore their expected behavior, accuracy, and use of alternative statistics determining reproducibility of such data [26, 27].

Here, we present a computational approach to generate synthetic ATAC-seq replicates to explore the behavior of various correlation and association metrics for epigenomics datasets. These synthetic ATAC-seq replicates are generated from eight true data sets to capture features uniquely present within ATAC-seq experiments. We have developed a random subsampling strategy to generate synthetic replicates with varying portions of shared peaks, as a proxy for reproducibility. Across our simulations, we apply the Pearson's R [37–39] and Spearman's ρ [40] and monitor their behavior, including the effect of removing co-zeros. Additionally, we demonstrate the behavior of other statistics, including non-parametrics such as Kendall's τ [41–44] and an information theoretic approach, normalized mutual information [45, 46], to determine their utility in assessing epigenomics data. Finally, we build a random forest model [47] using the normalized mutual information and R^2 coefficient between experiments to predict the biological relationships between replicates. Overall, our results demonstrate an improvement in the expected behavior of all statistics after removing co-zeros and normalized mutual information emerges as a promising statistic for measuring association between ATAC-seq samples.

Results

ATAC-seq data characteristics and subsamping strategy for synthetic replicate generation

To study the behavior of correlation measurements between ATAC-seq replicates (Fig. 1A), we analyzed data from three experiments using the A549, human lung cell line and implemented a subsampling paradigm to generate synthetic replicates. Across these experiments, the total number of reads mapped to the human reference genome varied from 15 million to nearly 43 million (Table 1). The number of genome-wide peaks found in the ATAC-seq samples varied across experiments and between replicates, ranging from approximately 80 to 130 thousand (Table 1). The fraction of sequenced read-pairs mapped in peaks (i.e. the FrIP score as defined by the ENCODE project [8, 11]), was greater than 0.34 for all of the A549 ATAC-seq samples (Table 1). These samples displayed high spatial correlation of peaks across replicates (Fig. 1B). Counting all whole fragments per kilobase per million (WFpkm), every ten kilobases, we observed a high statistical correlation between replicates, with average Pearson's R of 0.86, 0.87, and 0.94 (p-values < 0.05) between the technical replicates of the three biological replicate experiments (Fig. 1C).

For simulations, synthetic replicates were generated using the paired-end read alignment profiles from the eight ATAC-seq samples we generated. For each simulation, two synthetic replicates were initiated by duplicating a given true ATAC-seq experiment (Fig. 2A). Within the true ATAC-seq data set, reproducible, significant peaks were identified (see Methods). From these, a random portion of peaks was chosen to vary between the two synthetic replicates. This was accomplished by subsampling and removing 85% of the aligned sequenced fragments within each of the randomly chosen peaks between

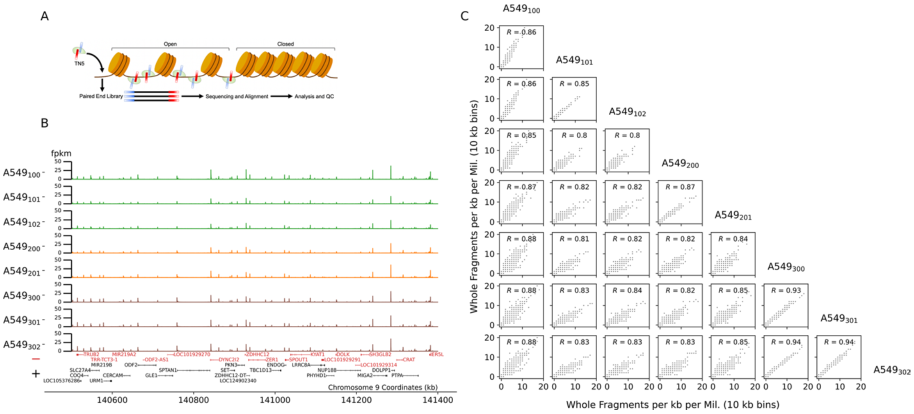

Fig. 1 ATAC-seq profiles of chromosome 9 form A549 cells. **A** TN5 binds to open chromatin, cutting DNA and adding primers to generate a paired-end sequencing library. **B** A549, ATAC-seq replicates along chromosome 9. Samples were generated using fresh cells (green) and previously cryo-preserved cell cultures (orange and brown). Positively (black) and negatively oriented genes are annotated along the bottom. **C** Pair-wise, bi-variate scatter plots of whole fragments per kb per million values (x- and y-axis) using 10 kb genomic bins between A549, ATAC-seq replicates. Sample names are annotated along the diagonal. Pair-wise Pearson's correlation statistic is annotated within subplots

Table 1 ATAC-seq experiments used, mapped reads, peak counts and FrIP scores

Sample title	Cell line	Mapped reads	MACS2 peaks	FrIP	Source
A549$_{000}$	A549	259,029,456	201,532	0.5898	ENCSR032RGS
A549$_{001}$	A549	329,679,445	194,975	0.5994	ENCSR032RGS
A549$_{002}$	A549	211,291,691	206,536	0.5596	ENCSR032RGS
A549$_{100}$	A549	23,987,725	110,323	0.588	This study
A549$_{101}$	A549	22,605,005	81,917	0.3404	This study
A549$_{102}$	A549	17,618,743	82,496	0.3702	This study
A549$_{200}$	A549	35,069,198	90,386	0.3515	This study
A549$_{201}$	A549	15,377,297	79,933	0.4202	This study
A549$_{300}$	A549	42,567,716	130,475	0.636	This study
A549$_{301}$	A549	28,744,542	107,737	0.6391	This study
A549$_{302}$	A549	35,836,016	117,087	0.6595	This study
GM12878$_{400}$	GM12878	46,889,870	114,746	0.7159	ENCSR095QNB
GM12878$_{401}$	GM12878	49,588,811	134,743	0.6452	ENCSR095QNB
HepG2$_{500}$	HepG2	48,113,686	173,756	0.4257	ENCSR042AWH
HepG2$_{501}$	HepG2	48246,610	135,767	0.4605	ENCSR042AWH
IMR-90$_{600}$	IMR-90	47,543,633	178,156	0.5363	ENCSR200OML
IMR-90$_{601}$	IMR-90	61,359,070	200,216	0.6104	ENCSR200OML
K562$_{700}$	K562	48,217,636	178,230	0.5112	ENCSR483RKN
K562$_{701}$	K562	52,270,533	176,789	0.5196	ENCSR483RKN
RWPE2$_{800}$	RWPE2	55,152,003	166,239	0.474	ENCSR080SNF
RWPE2$_{801}$	RWPE2	43,166,947	177,496	0.4555	ENCSR080SNF
RWPE2$_{802}$	RWPE2	48,162,285	154,758	0.4652	ENCSR080SNF
WTC11$_{900}$	WTC11	74,558,506	245,677	0.5505	ENCSR541KFY
WTC11$_{901}$	WTC11	79,335,328	277,824	0.5732	ENCSR541KFY

the two synthetic replicates (Fig. 2B, C). This process was repeated fifteen times for each of the eight real ATAC-seq samples, randomly varying the common peaks from 5 to 99% of peaks between the two synthetic replicates. Finally, across all 120 simulations, for each pair of synthetic replicates, the WFpkm values were calculated in ten kilobase windows and used in statistical comparisons (Fig. 3A).

Top–down correlation displays best behavior in correlation analysis across simulations

Across the 120 down sampling simulations, correlation and association statistics were calculated between each pair of synthetic replicates. The Wfpkm counts were used between synthetic replicates in statistical analysis (Fig. 3A). The values of correlation (Fig. 3B) and association (Fig. 3C) statistics were calculated within each simulation as a function of the number of shared peaks between synthetic replicates. For each examined statistic, the area under the curve (AUC), formed by the statistical values calculated across portions of shared peaks, for each simulation was used in comparisons (Additional file 1: Fig. S1). Of the correlation statistics, the Top–Down correlation statistic had the smallest average AUC of 0.6881 (95% CI 0.6860–0.6906) and was significantly smaller than the average AUC of the Pearson's R, at 0.8284 (95% CI 0.8237–0.8335, p-value $= 0$, bootstrapped difference of mean AUC). Both the two non-parametric statistics, Spearman's ρ and Kendall τ, had significantly larger average AUCs compared against the Pearson's R (p-values $= 0$, bootstrapped difference of

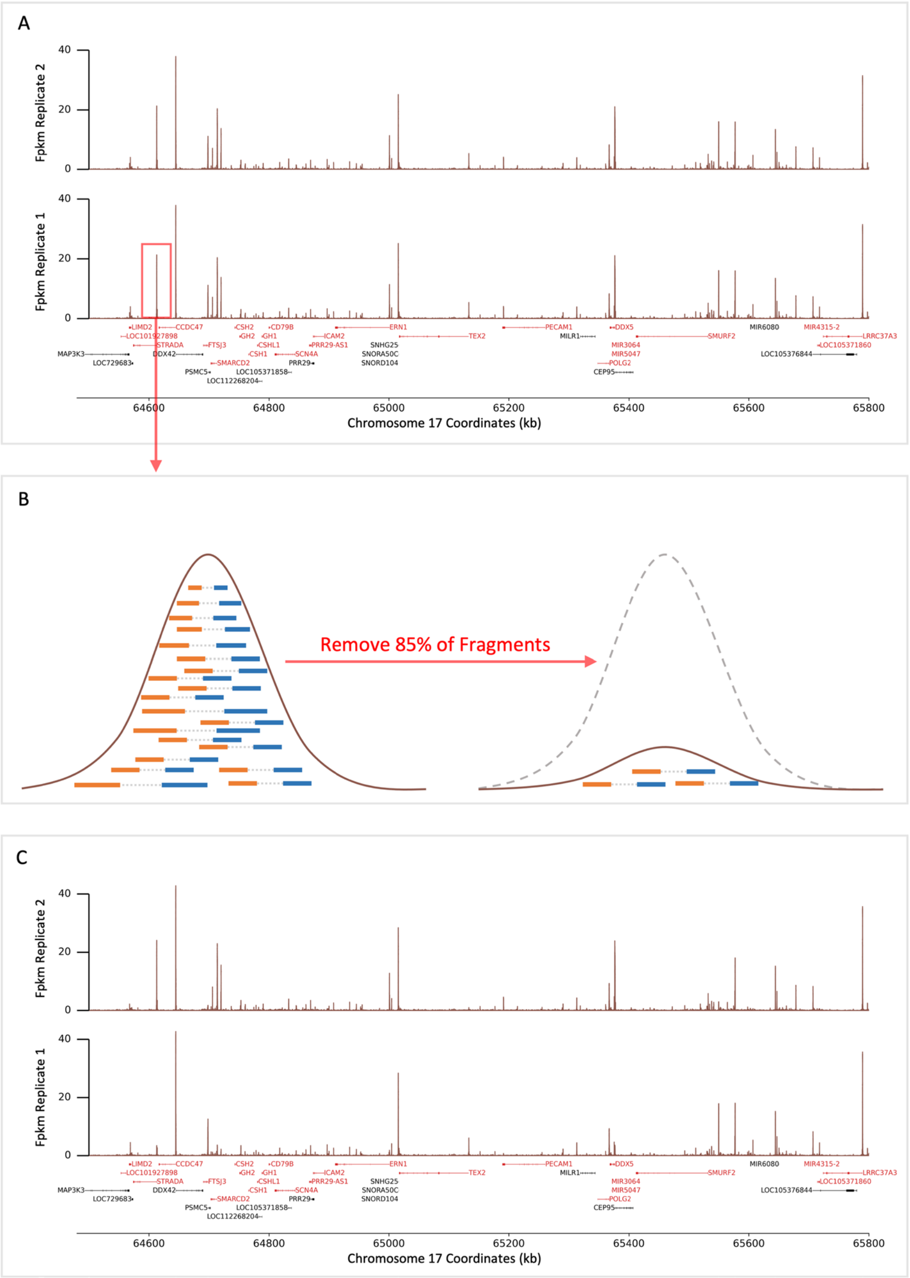

Fig. 2 Synthetic replicate generation via peak down-sampling. **A** An example region along chromosome 17 of true, A549 ATAC-seq data. Real ATAC-seq signal (brown lines) is used to initialize two synthetic replicates. Red and black horizontal bodies depict negatively and positively oriented genes, respectively. **B** A portion of the genome-wide significant peaks (ranging from 0 to 1) are chosen randomly between the two synthetic replicates. Within one of the replicates, 85% of paired reads (blue and orange rectangles connected by grey dotted line) are removed to down-sample signal within that locus. **C** Example of two synthetic replicates with a known portion of peaks varying between them

mean AUC). However, they demonstrated nearly identical AUC profiles compared to each other, with average AUC of 0.9140 (95% CI 0.9118–0.9162) and 0.9096 (95% CI 0.9074–0.9120) respectively (p-value = 0.037, bootstrapped difference of mean AUC).

Across the metrics of association, Kendall's W, normalized mutual information, and the R^2 coefficient, between replicates, the R^2 coefficient exhibited the greatest sensitivity to the change in portion of shared peaks between synthetic replicates (Fig. 3C). Across simulations, the average AUC of the R^2 coefficient was 0.7026 (95% CI 0.6951–0.7102). This average AUC was significantly smaller than the average AUC of the Kendall's W and normalized mutual information, with values of 0.957 (95% CI 0.9559–0.9581) and 0.8197 (95% CI 0.8153–0.8241), respectively (p-value = 0, bootstrapped difference of mean AUC).

Removal of co-zeros improves estimates of correlation and associations

Using this simulation paradigm, we evaluated the efficacy of removing co-zeros from the analysis to determine the impact on correlation and association statistics. Co-zero values were defined as value counts in ATAC-seq experiments that appeared to have zero aligned fragments in a genomic bin of ten kilobases between two replicates (Additional file 2: Fig. S2). On average, these values can make up nearly 5% of a given bi-variate distribution formed between real ATAC-seq replicates (Additional file 3: Fig. S3). Across all the correlation and association statistics examined here—except for Top–Down correlation—removing the co-zero values significantly reduced the average AUC (Table 2, Fig. 3B, C, Additional file 1: Fig. S1). The large reduction observed in the AUC after removing co-zeros from analysis was unexpected, as co-zeros are a modest portion of the bi-variate distribution formed between two replicates.

After removing co-zeros, all the correlation statistics, Top–Down correlation, Pearson's R, Spearman's ρ, and Kendall's τ, displayed nearly identical sensitivity to the change in shared peaks between replicates across simulations (Fig. 3B). However, the Pearson's R had the largest average AUC of 0.6965 (95% CI 0.6946–0.6984) followed by the Top–Down statistic (AUC of 0.6872, 95% CI 0.685–0.6895, p-value = 0, bootstrapped difference of mean AUC). The Spearman's ρ (mean AUC: 0.6686, 95% CI 0.6665–0.6705) and Kendall's τ (mean AUC: 0.6673, 95% CI 0.6654–0.6691) statistics had the smallest and identical average AUC after removing co-zeros (p-value = 0.208, bootstrapped difference of mean AUC). Furthermore, the AUC of the Top–Down correlation statistic was unaltered by the exclusion of co-zero values between synthetic replicates (Fig. 3B, Additional file 1: Fig. S1, Table 2, p-value = 0.635, bootstrapped difference of mean AUC). This observation was not surprising given how Top–Down correlation places emphasis on larger values, down-weighting smaller values, such as co-zeros [48].

(See figure on next page.)

Fig. 3 Synthetic replicate bivariate plots and statistical profiles. **A** Scatter plots displaying counts per genomic bin (10 kb in size) of whole fragments per kilobase per million (WFpkm) between two synthetic replicates (x- and y-axis) generated in process Fig. 2A–C. The percentage of shared peaks decreases between the two simulated replicates from top to bottom. **B** Correlation values (y-axis) as a function of percentage of shared peaks between synthetic replicates (x-axis). **C** Association scores (y-axis) as a function of the percent of shared peaks between synthetic replicates (x-axis). In **B** and **C**, red and grey curves depict the mean and 95% confidence interval (respectively) across simulations. A dashed line marks a one-to-one relationship between the x- and y-axis. Left and right columns display change in values as a function of removing co-zeros

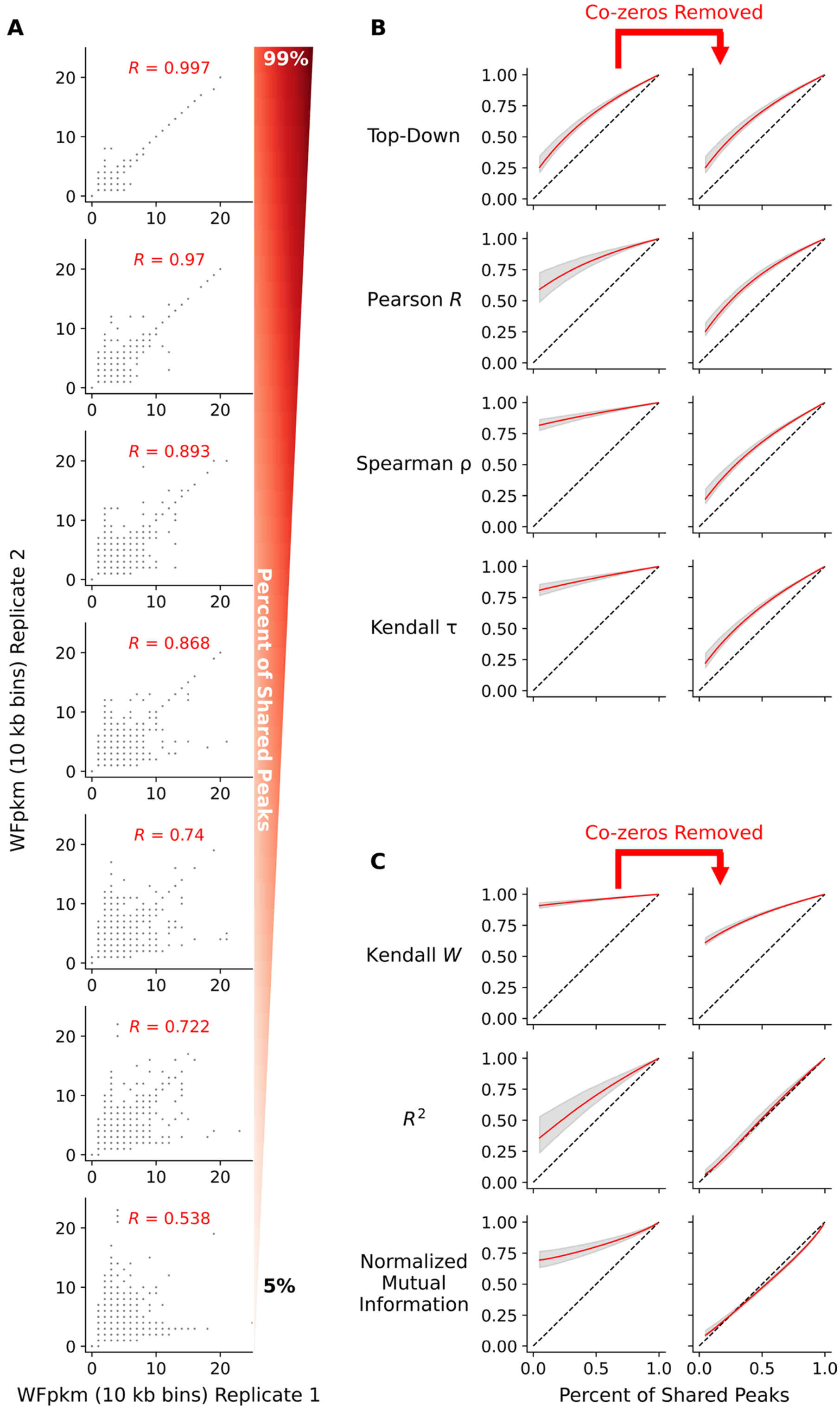

Fig. 3 (See legend on previous page.)

Table 2 Mean area under the curve across simulations

Statistic	Mean (95% CI)	Mean (95% CI)—Co-zeros removed	*p*-value[a]	σ^{2b}
Top–down correlation	0.6881 (0.6860–0.6906)	0.6872 (0.6850–0.6895)	0.635	–
Pearson *R*	0.8284 (0.8237–0.8335)	0.6965 (0.6946–0.6984)	0.0	0.0201
R^2	0.7026 (0.6951–0.7102)	0.5346 (0.5324–0.5368)	0.0	0.0329
Spearman ρ	0.9140 (0.9118–0.9162)	0.6686 (0.6665–0.6705)	0.0	0.0136
Kendall τ	0.9096 (0.9074–0.9120)	0.6673 (0.6654–0.6691)	0.0	–
Kendall W	0.9570 (0.9559–0.9581)	0.8343 (0.8333–0.8353)	0.0	–
Normalized Mutual Information	0.8197 (0.8153–0.8241)	0.5055 (0.5045–0.5065)	0.0	0.016

[a] The *p*-value represents the test of differences in mean AUC after removal of co-zeros

[b] Variation values were calculated during analysis of data from true ATAC-seq experiments

Normalized mutual information best estimates difference between replicates

Removing co-zero values had a similar effect on association metrics, attenuating and improving the average AUC across the portion of shared peaks between synthetic replicates (Fig. 3C, Additional file 1: Fig. S1). Apart from Kendall's W, the R^2 coefficient and normalized mutual information, on average, displayed a nearly one-to-one relationship with the portion of shared peaks between replicates (Fig. 3C). The average AUC of normalized mutual information was 0.5055 (95% CI 0.5045–0.5065) and was smaller than the average AUC of the R^2 coefficient, with a value of 0.5346 (95% CI 0.5324–0.5368, *p*-value $= 0$, bootstrapped difference of mean AUC). This difference in average AUC indicates that normalized mutual information better follows the designed proportion of shared peaks between synthetic replicates across our simulations, compared to the R^2 coefficient.

As introduced earlier, one parameter in this simulation is the removal of a percentage of aligned read-pairs from within randomly selected peaks (Fig. 2B). Initially set at 85%, this parameter was altered to simulate ATAC-seq replicates that are nearly reproducible (at 50%) at every selected peak or broadly unreproducible (at 95%) across all selected peaks. Comparing the results between the two simulation sets with 85 and 95% of reads removed, we observed no significant difference between the two simulations (Additional file 4: Fig. S4, Additional file 5: Fig. S5). This is somewhat expected when considering the small difference in magnitude between removing 85 and 95% of reads from within peaks. In simulations with only 50% of read pairs removed from selected peaks, after removing co-zeros, the two statistics that showed the largest response in our simulation were the R^2 coefficient and normalized mutual information.

Co-zeros inflate estimates of correlation and association in epigenomic assays

After successfully implementing normalized mutual information between simulated replicates, we next examined how this statistic behaves when used on replicates from real epigenomic experiments. We also monitored how dropping co-zeros affects estimates of correlation (and association) between samples. For these analyses, additional experiments were downloaded from the ENCODE project public repository [11]. These included additional ATAC-seq experiments (Table 1) as well as ChIP-seq experiments, specifically twenty and ten assays for the H3K27ac and H3K4me3 modifications, respectively (Additional file 10: Table S2). The Spearman's ρ, Pearson's R, R^2 coefficient, and normalized mutual information were calculated between these replicates. Correlation and association values were also calculated between non-replicates within each of the three assays. We then repeated these comparisons, eliminating co-zeros from calculations. With this design, we were able to gauge the effect of masking co-zeros within replicates, between non-replicates (within the same assay), and across different types of epigenomic data. Between real experiments, excluding co-zeros from analysis significantly decreased the computed correlation and association statistics (p-value $< 1^{-10}$, Wilcoxon signed-rank test). This reduction is seen in the distributions of the Spearman's ρ, Pearson's R, R^2 coefficient, and normalized mutual information (Additional file 6: Fig. S6) across all three assays, ATAC-seq (Fig. 4A), H3K27ac (Fig. 4B), and H3K4me3 (Fig. 4C). Further investigation revealed that omitting co-zeros primarily alters estimates of correlation and association between non-replicates (black lines and dots in Fig. 4), which were significantly decreased (p-value $< 1^{-19}$, Wilcoxon signed-rank tests). The correlation and association values between true replicates from H3K27ac and H3k4me3 assays were unaltered when ignoring co-zeros (p-value > 0.001, Wilcoxon signed-rank tests). However, a significant alteration in correlation (and association) estimates between replicate ATAC-seq experiments (red lines and dots in Fig. 4A) was detected

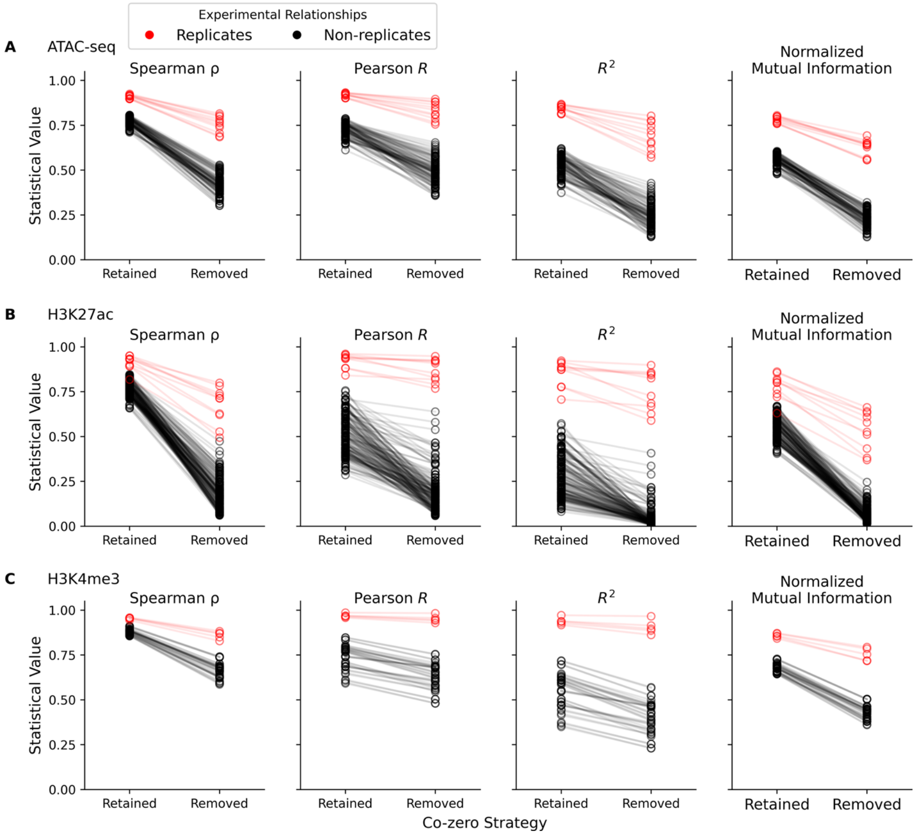

Fig. 4 Correlation and association statistics across epigenomic experiments. For samples from (**A**) ATAC-seq and ChIP-seq (assays for (**B**) H3K27ac and (**C**) H3K4me3 modifications) experiments, the Spearman's ρ, Pearson's R, R^2 coefficient, and normalized mutual information (y-axis of columns left to right, respectively) were calculated on WFpkm counts between replicates, with and without co-zeros (x-axis). Red and black dumbbells represent calculations between replicates or non-replicates, respectively, and connect calculations across the co-zero handling strategy

(p-values < 0.0009, Wilcoxon signed-rank tests). Moreover, excising co-zeros expanded the difference in the average estimates of correlation and association between replicates and non-replicates, across all three assays (Fig. 4, Additional file 11: Table S3). Thus, removing co-zeros produced lower estimates of correlation and association for most samples, and overall improves the ability to differentiate pairs of replicates from non-replicates in real epigenomic data.

A random forest prioritizes mutual information for predicting replicate relationships

After removing co-zeros, the R^2 and normalized mutual information metrics performed best in simulation. Furthermore, when tested on real ATAC-seq experiments, these two metrics produced the largest difference (on average) between replicates and non-replicates. Given the comparable behavior of normalized mutual information and the R^2 coefficient on true ATAC-seq data we set out to further assess their usefulness in predicting the relationships between experiments. To do this we combined our A549 ATAC-seq samples with ATAC-seq samples from the ENCODE project. These included additional biological replicates of the A549 cell line, as well as ATAC-seq experiments in the HepG2, RWPE2, GM12878, IMR-90, K562, and WTC11 cell lines (Table 1). With these combined data, comparisons between any two ATAC-seq experiments were classified into one of three discrete, replicate classes; (1) between independent ATAC-seq experiments in different cell lines, (2) between independent experiments using the same cell line, or (3) between true replicates. Plotting the normalized mutual information and R^2 coefficient calculated between ATAC-seq experiments with the above classifications revealed clustering of the classes between replicates (Fig. 5A). Between the two statistics, the R^2 coefficient displayed the largest variation across compared experiments (Table 2) and the most mixing of the three class labels (Fig. 5A). We also observed a strong co-linear relationship between the calculated R^2 coefficient and normalized mutual information scores (Additional file 7: Fig. S7, Pearson's $R = 0.96$, p-value $< 1^{-10}$).

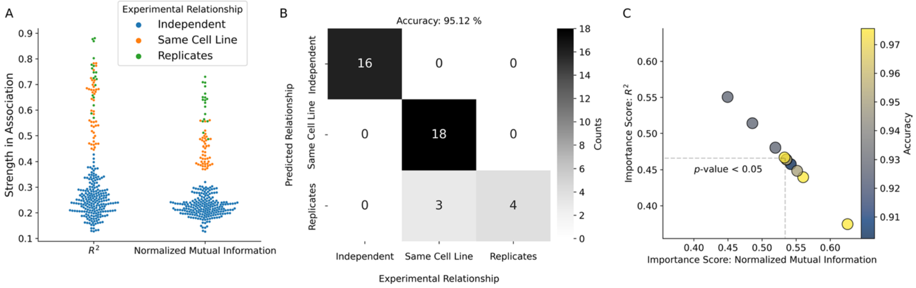

Fig. 5 Random forest prediction of experimental relationships. **A** Distributions of the coefficient of determination (R^2) and normalized mutual information scores calculated on binned counts of WFpkm between ATAC-seq experiments. Blue, orange, and green dots mark comparisons between independent experiments, independent experiments using the same cell line, and true experimental replicates, respectively. **B** Example confusion matrix from a random forest model using R^2 and normalized mutual information as features to predict experimental relationships (y-axis) presented in **A** (x-axis). The confusion matrix depicts results of model on a hold-out set (40% of data, accuracy = 95.12%). Light to dark colors depict the number of counts per class. **C** Bi-variate plot displaying the change of paired importance scores from ten-fold cross validation between the normalized mutual information (x-axis) and R^2 (y-axis) features. Dashed lines depict the uni-variate means of the normalized mutual information and R^2 scores. Blue and yellow colors depict the level of accuracy for each fold

This finding is not surprising given that both metrics attempt to measure the same relationship between samples.

To quantify which statistic (the R^2 coefficient or normalized mutual information) better estimates experimental relationships between replicates, we built a random forest model. This model uses the reported values of the R^2 coefficient and normalized mutual information between ATAC-seq experiments as features to predict the replicate class (as defined above). We utilized ten-fold cross validation, stratifying on the replicate class to build our random forest. An example confusion matrix from one of these folds demonstrates the model had difficulty distinguishing between independent experiments using the same cell line and true experimental replicates (Fig. 5B). This difficulty also manifested as lower f1-scores and recall for this class (Additional file 8: Fig. S8). The accuracy across these folds ranged from 88 to 98% (Fig. 5C). Across the folds, the feature importance score of the R^2 coefficient was inverted with that of normalized mutual information (Fig. 5C). Overall, we observed a greater feature importance score for normalized mutual information, with a significant average pair-wise difference between the R^2 coefficient and normalized mutual information of 6.78% (p-value < 0.05, Wilcoxon signed-rank test).

Discussion

To improve the assessment of reproducibility in epigenomic data sets, we sought to investigate the use of several correlation and association statistics on binned genomic signals. Our findings suggest that best practices should include analyzing association between compared replicates (or experiments) via normalized mutual information with binned, Fpkm counts rounded to the nearest whole integer, after the removal of co-zero values as input. In choosing a correlation statistic, after removing co-zero values, our results indicate little difference in the outputs from the Pearson's R, Spearman's ρ, Kendall's τ, or Top–Down correlation statistics. Notably, from simulations, we observed that the Top–Down correlation statistic was unaffected by the removal of co-zeros values. As such, this statistic should serve as an alternative for investigators if binned co-zeros values between replicates are retained.

As part of this study, we generated highly correlated, new ATAC-seq experimental replicates of the A549 cell line. Our data highly correlates with previously published ATAC-seq data of the A549 cell line generated by the ENCODE project. Using these data, we generated a novel simulation that utilizes down sampling to generate replicates with known varying signals. While similar simulation studies have been conducted on Hi-C sequencing data [30], to our knowledge, no prior study has examined the behavior of statistical metrics on ATAC-seq data. That said, there are several statistics and methodologies that may be used to analyze this data type, such as Poisson regression [49]. Improving on this simulation design could help generate a framework that allows researchers to develop new statistical tools for hypothesis testing.

In our simulations, we observed that most statistics overestimate the correlation of signal between replicates. One specific strategy we investigated to reduce this inflation was the removal of co-zeros, which is an option present in several bioinformatic software suites [29, 31, 34]. Our analysis demonstrated that removal of these values can provide a more accurate estimate of correlation between replicates as measured by the known number of peaks between replicates. Interestingly, we never observed a correlation value that perfectly trends with the designed number of peaks between synthetic replicates. We also did not observer negative correlation values between the replicate Fpkm counts. The first of these observations can be explained by background autocorrelation still present within our synthetic replicates. The second of these observations may point to a limitation in the design of our simulation, as negative correlation values have been observed in true ATAC-seq profiles [20, 31].

In epigenomics and chromatin accessibility data sets, biological interpretation of the data is dependent upon visualization of "peaks" where accumulation of sequenced reads denotes the presence of a modification or an accessible region. Regions with zero (or nearly zero) aligned sequenced reads are deemed unmodified or inaccessible and largely ignored when interpreting data. Correlation statistics should provide biologists with the confidence that replicates are truly comparable. As stated above, the inclusion of co-zeros seems to inflate values of most correlation and association statistics. Thus, removal of co-zeros formed by the genomic bins that overlap and account for inaccessible regions may be warranted.

Using our simulation, we also examined the behavior of three association statistics, which we distinguish from the set of correlation statistics as those metrics ranging in value from zero to one. These association statistics were the R^2 coefficient, normalized mutual information statistic, and Kendall's W. Prior to the removal of co-zeros, the only association statistic that displayed any sensitivity to the change in shared peaks between replicates was the R^2 coefficient. Co-zeros inflate the value of this statistic by reducing the total summed error between data points during calculation. Similarly, co-zeros increase the information gained between replicates when calculating the normalized mutual information score. In other words, knowing a replicate has a value of zero at a given genomic bin provides information that there is a zero at the corresponding bin within the other replicate. After removing co-zeros, we saw a large improvement in the sensitivity of both these statistics.

Curiously, Kendall's W displayed the least sensitivity to the designed peak counts between synthetic replicates. This statistic was of interest given Kendall's W is capable of simultaneously examining the ranks of more than two input samples [41, 50]. This would have provided researchers with a statistical tool capable of examining correlation among a full set (triplicate) of replicates within a single test, rather than multiple pairwise comparisons. Removing co-zeros did little to improve the sensitivity of this statistic. The other statistic from Kendall, Kendall's τ, displayed similar performance to the other non-parametric statistic, Spearman's ρ. This finding is contrary to other studies of Kendall's τ conducted in the fields of signal processing and psychology [43, 44]. For analysis of genomic data, the Spearman's ρ is standard in deeptools' correlation functions [13].

We also examined the effect of dropping co-zeros when estimating correlation between real ATAC-seq and ChIP-seq samples. Much like our results from analysis on simulated ATAC-seq replicates, expunging co-zero values from correlation (and association) calculations reduced the value of the reported statistics between real samples. These effects were primarily seen in the correlation and association scores calculated between non-replicates within the examined assays. In particular, the correlation and association values between true H3K27ac and H3K4me3 replicates were unaffected by eliminating co-zeros. These correlation scores were high and remained high after excising co-zeros from calculations. This may be due to higher overall signal in these assays. Importantly, omitting co-zeros from analysis produced a larger difference in the average correlation between groups of replicates and non-replicates. Thus, excluding co-zeros from analysis is an important step for quality control procedures looking to identify errant samples.

Of the statistics examined here, the R^2 coefficient and normalized mutual information score were the most sensitive to the change in shared peaks between replicates (when co-zeros were removed). Comparison of these two statistics revealed that normalized mutual information was the better-behaved statistic. This behavior manifested as smaller AUC within simulations, less variation in values, and larger differences in values between groups of replicates and non-replicates. Similarly, the computational evidence provided by our random forest model suggests that normalized mutual information was better at estimating experimental relationships between true ATAC-seq replicates. Taken together, these results indicate that of the two metrics, normalized mutual information may be the stronger association metric for ATAC-seq data. Information theoretic approaches, such as normalized mutual information, have been utilized in several other biological fields, ranging from cancer genomics to fungal genetics [51–57]. Regarding ATAC-seq data, a handful of other studies have specifically used mutual information in data integration, analysis, and deep-learning of single-cell ATAC-seq profiles [58, 59]. For those investigator interested in using information theoretic approaches, several of these functions are made easily available within the python, scikit learn library [46].

To perform correlation and association analyses as seen here, we have generated python code and an executable for public use. These software, installation instructions, and a tutorial written as a jupyter notebook are hosted on the Github listed within the data availability section. We hope these tools will benefit investigators and students in their exploration of the mutual information statistic and the effect of excluding co-zero values in their epigenetic data.

Sparsity and zero mapped sequenced reads are not unique properties of ATAC-seq data. These extend to genomic, Hi-C, ChIP-seq, and RNA-seq data sets. Imputation along with modified zero-inflated models have been used with success for studying RNA sequencing data sets and detecting regions with differential expression [60]. Simulations and models of sampling zero-genomic count data have been developed to understand the effects of these values, particularly in the context of differential analysis [36]. Previous simulation studies of ATAC-seq have been focused on generating ATAC-seq data, for pipeline development, or single-cell ATAC-seq samples, to examined different approaches in their analysis [61, 62]. To our knowledge, this is the first example of using a simulation approach for studying reproducibility and association of ATAC-seq samples. Adapting strategies from these previous studies will help improve our simulation and expand it to other genomic and epigenomic sequencing data. The current results of our study strongly suggest that normalized mutual information is an appropriate metric for measuring reproducibility in chromatin accessibility assays.

Conclusions

For this study, we produced eight ATAC-seq experiments using the A549 Cancer cell line. Across replicates, these ATAC-seq samples are well correlated and reproducible. For investigations of chromatin accessibility (particularly in the A549 cell line), these experiments are an additional resource for developing analysis pipelines, peak detection algorithms, and machine learning approaches.

Leveraging the A549 ATAC-seq experiments, we designed a computational simulation to generate simulated replicates. Specifically, synthetic replicates were coded that share a known, fixed portion of significantly enriched loci. Using these replicates, correlation metrics—the Pearson's R, Spearman's ρ, Top–Down, and Kendall's τ—and association statistics (ranging from zero to one)—the R^2 coefficient, Kendall's W, and normalized mutual information—were tested for accuracy. Overall, the reported value of these statistics was inflated and much larger than the fixed portion of shared, significant loci between replicates.

Removing specific outliers from ATAC-seq data, specifically the removal of co-zeros, improved estimates of correlation and association. We estimate that co-zero values, when comparing WFpkm counts between two real ATAC-seq experiments, occupy nearly five percent of a bi-variate distribution. While only a small portion of the total data, filtering these values from analysis greatly improves the measurements of most correlation and association statistics between samples, in simulation. Applied to real ATAC-seq and ChIP-seq data, removing co-zero values from comparison significantly reduced the reported correlation and association statistic, matching results from simulation.

One of the association statistics examined here is normalized mutual information, an information theoretic approach that is less well known across the (epi)genomics field. After removing co-zero values, normalized mutual information displayed the lowest inflation relative to the similarity between simulated replicates. The R^2 coefficient also performed well in simulations (after removal of co-zeros), displaying good sensitivity to differences between simulated replicates. Of these two association metrics, a random forest model selected normalized mutual information as the stronger feature when estimating experimental relationships between real ATAC-seq experiments. From these results we conclude that normalized mutual information is a powerful, non-parametric approach for estimating association between ATAC-seq experiments.

Methods

Construction of A549 ATAC-seq libraries

ATAC-seq experimental libraries were generated using A549 human lung carcinoma epithelial cells (ATCC, VA, catalog #CCL-185) [63–65]. Three biological replicate libraries were prepared from freshly harvested cells using an ATAC-seq kit (Active Motif, 53150) following the manufacturer's protocol. The remaining five libraries were prepared using cryopreserved cells following methods outlined in Milani et al. with modifications [18]. Briefly, A549 cells were cultured in T75 flasks and harvested by trypsinization using 0.25% (w/v) Trypsin-EDTA (0.5%) solution (Gibco, 15400054). Harvested cells were centrifuged and pellets resuspended in freezing media containing DMEM (Gibco, 11885-084), 10% FBS (Corning, 35-015-CV), and 10% DMSO (ATCC, 4-X). Pellets were frozen using an isopropyl alcohol chamber (Thermo Fisher Scientific, 5100-0001) at − 80 °C. After 24 h, frozen cells were transferred to liquid nitrogen for long term storage. To perform experiments, cryopreserved cells were transferred to − 80 °C for several days, and the tube was immersed in 37°C water bath for approximately two minutes on the day libraries were prepared. Thawed cells were resuspended in 1X PBS with protease inhibitor cocktail (Thermo Fisher Scientific, 78430). Cell counts and viability were assessed and aliquots containing 80,000 cells per sample were processed into ATAC-seq libraries.

Sequencing, alignment and filtering

ATAC-seq libraries were sequenced at the sequencing facility at Los Alamos National Laboratory on an Illumina NextSeq2000 sequencer in paired end mode (PE151) using P3 chemistry. With Fastp, raw reads were trimmed and filtered to remove Nextra adaptors and reads with repetitive sequences [66]. Additionally reads were also filtered to remove bases with low quality scores (`q` < 15). These processed reads were aligned to the new, telomere-to-telomere human reference genome, version 2 [67] via bwa [68]. After alignment, duplicate sequenced pairs were marked via samblaster and removed from analysis [69]. Read pairs mapping to the mitochondria were also removed (see Additional file 9: Table S1).

Other data used

Raw ATAC-seq data, in the form of paired fastq.gz files, was downloaded from the ENCODE project for the A549, HepG2, RWPE2, GM12878, IMR-90, K562, and WTC11 cell lines [11, 70]. The ENCODE file experiment and replicate accession numbers are included in Table 1. For alignment, these data were passed through the same pipeline described above for ATAC-seq samples generated here, and aligned to the human, telomere-to-telomere, reference genome [67].

For ChIP-seq experiments, twenty and ten assays of the H3K27ac and H3K4me3 epigenetic modifications (respectively) were downloaded (also) from the ENCODE project as raw alignments in bam file format. The ENCODE accession numbers of these files are listed in Additional file 10: Table S2. Each of these raw alignments were made with the GRCh38 (v1.5.1) human reference genome. Filtered bam files were generated via `samtools view` command and the following flags: `-F 4 -F 256 -F 512 -F 1024 -F 2048 -q 30`. When comparing differences between correlations and association values within experiments, between calculations with and without co-zeros, and between groups of replicates and non-replicates, a Bonferroni correction was used to calculate the adjusted p-value $= 0.05/(12 \times 3) \sim 0.00139$, for establishing significance.

Peak calling, peak filtering and reproducibility

After filtering, sample alignments were analyzed to identify loci displaying significant enrichment of paired-end reads. This peak calling was conducted using MACS2 [6, 71]. Specifically, after removing duplicates and mitochondrial mapped reads, samples were further filtered using samtools with the following flags:

`-F 4 -F 256 -F 512 -F 1024 -F 2048 -q 30` and then passed to MACS2 in `BAMPE` mode [72, 73].

Between true, biological replicates, reproducible peaks were identified via irreproducible discovery rate thresholding [74]. Using ChIP-R, replicate narrow peak files were filtered to retain only those peaks that were consistent across all replicates; in ChIP-R, where command line parameter, `m` = number of biological replicates [75]. In addition to this setting the '`-fragment`' option was also invoked. These sets of final peak counts were retained for further analysis.

Genomic down-sampling and simulation design

For each of the eight ATAC-seq experiments of A549 cells generated in this study, synthetic replicates were generated by duplicating a given sample into two copies and then randomly, varying the total number of shared peaks between them. Specifically, for a given ATAC-seq experiment, a set portion of peaks was chosen at random, such that within one of the synthetic replicates, a given selected peak was depleted, randomly removing a portion of the alignments within the peak bounds (as defined by MACS2). These sets of peaks were randomly selected from the set of reproducible peaks for that sample and its associated biological replicates (see above). Three sets of simulations were conducted, removing 50, 85 and 95% of reads within selected peaks. This procedure results in two synthetic ATAC-seq replicates, generated from a single, true parent ATAC-seq data set. These synthetic 'sister' ATAC-seq data sets have identical genome-wide alignments except within a sub-set of loci that vary between them. From each true ATAC-seq data set, synthetic sister replicates were generated by varying the total percentage of shared peaks from 99 to 5%, with a delta of 5%. For each simulation, across the change in portion of shared peaks, a common random seed was used to preserve autocorrelation across this axis. This process was repeated fifteen times for each of the eight, A549 ATAC-seq samples, totaling a one hundred and twenty simulations.

Genomic binning, fragment counts, and standardization

On both synthetic samples from simulation studies or replicates from (true) ATAC-seq experiments, a genomic binning approach was used to estimate correlation and association statistics between samples. For each chromosome, contiguous bins were established 5'–3', every ten kilobases. Within each of these bins, the number of sequenced fragments is counted and standardized to fragments per kilobase per million. These counts were rounded up to their nearest whole integer generating standardized counts of whole fragments per kilobase per million (WFpkm).

Calculating correlation and association metrics

In python scripts, using the scipy-stats module [76], the Pearson's R, Spearman's ρ, and Kendall's τ were calculated on the WFpkm counts between pairs of ATAC-seq replicates. Functions for the Top–Down correlation metric [48] and Kendall's W rank statistic [41, 50] were also developed using custom python code. The R^2 coefficient was calculated using the square of the Pearson's R. The normalized mutual information statistic from pythons sklearn module [46] was used in association studies. Between any pair of WFpkm counts, the bi-variate distribution was examined to identify instances were both profiles contained a value of zero WFpkm. For studies of the effects of co-zero inflation, these co-zero values were removed, and the correlation (or association) statistics recalculated on these filtered distributions.

For correlation analysis on ATAC-seq experiments conducted here using A549 cells, the Pearson's R correlation statistic was calculated on WFpkm values between replicates with co-zeros removed. Similarly, co-zeros were removed prior to calculating correlation and association statistics between replicates of ATAC-seq data downloaded from the ENCODE project public repository.

Statistical tests on area under the curve

Across simulations, values of correlation and associations statistics were calculated as a function of the designed portion of peaks between synthetic replicates. For each statistic tested, the 95% confidence interval of the average area under the curve was calculated via bootstrapping, with a thousand iterations. This was done for statistical profiles from simulations with and without co-zero values. For comparisons of the average area under the curve, either between statistics or within the same statistic after removing co-zeros, one thousand permutations were used to calculate the null distribution of the difference between the mean area under the curve [77]. The proportion of these differences greater than or equal to the true observed difference was used as the p-value. A significance level of 0.05 was used to reject the null hypothesis, H_0: no difference in mean area under the curve, in favor of our alternative hypothesis, H_1: difference of mean area under the curve.

Design of random forest model

A random forest model was built in python using the scikit learn module [46, 47]. Association statistics from the ATAC-seq data generated in this study on A549 cells and additional ATAC-seq data downloaded from the ENCODE project was used as input (see Table 1). As features in this random forest, the R^2 coefficient and normalized mutual information were calculated between every pair of ATAC-seq experiments using WFpkm counts, across ten kilobase pair, genomic bins and removing co-zero values. The comparison of each unique pair of experiments (totaling 276) were discretized as (1) between independent ATAC-seq experiments in different cell lines, (2) independent experiments using the same cell line, and (3) between true replicates. The total number of comparisons distributed among these three classes was 213, 45, and 18 (respectively). Given the over-representation of comparison between independent ATAC-seq experiments in different cell lines, 39 of the 213 comparisons were chosen randomly to represent the total, unique comparisons between experiments with unique cell lines. This down sampling resulted in 39, 45, and 18 comparisons between independent experiments in different cell lines, independent experiments using the same cell line, and true replicate experiments, respectively.

For the testing and training of the model, test and training sets of the classes defined above were selected using a stratified, 40:60 split of the data. Additionally, ten-fold, stratified cross validation was used to train and test the model [78]. A hundred estimators with the entropy selection criterion were used along with default settings in the python random forest classifier function within scikit learn [46].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05553-0.

Additional file 1: Figure S1. Boxplots displaying the area under the curve (y-axis) across statistics (x-axis) with co-zeros retained and removed from analysis (blue andorange boxes, respectively).

Additional file 2: Figure S2. Bi-variate plot of WFpkm counts (across 10 kb genomic bins) between replicates of real, A549ATAC-seq experiments. Dark red to blue colors and marker size designate the density (log10 (WFpkmcounts)) of counts between replicates. Co-zero values appear as an orange dot in lower left corner. A dashedgrey line represents a one-to-one relationship between the two replicates.

Additional file 3: Figure S3. The percent of co-zero values in bi-variate WFpkm distributions between real ATAC-seq experiments.Sample names are annotated along the x- and y-axis.

Additional file 4: Figure S4. Correlation and association values (y-axis) as a function of percentage of shared peaks betweensynthetic replicates (x-axis). Red and grey curves depict the mean and 95% CI (respectively) values across-simulations. A grey, dashed line marks a one-to-one relationship between the x- and y-axis. Left and rightcolumns display change in values as a function of removing co-zeros. Results are from simulations with 50%paired reads within selected peaks removed.

Additional file 5: Figure S5. Correlation and association values (y-axis) as a function of percentage of shared peaks betweensynthetic replicates (x-axis). Red and grey curves depict the mean and 95% CI (respectively) values across-simulations. A grey, dashed line marks a one-to-one relationship between the x- and y-axis. Left and rightcolumns display change in values as a function of removing co-zeros. Results are from simulations with 95%paired reads within selected peaks removed.

Additional file 6: Figure S6. Correlation and association statistics across epigenomic experiments. For samples from **A**ATAC-seq and ChIP-seq (assays for **B** H3K27ac and **C** H3K4me3 modifications) experiments, the Spearman'sρ, Pearson's R, R^2 coefficient, and normalized mutual information (x-axis of columns left to right, respectively) were calculated on WFpkm counts between replicates, with (blue) and without co-zeros (orange).

Additional file 7: Figure S7. The coefficient of determination (R^2) versus the normalized mutual information (y- and x-axis,respectively) calculated on binned counts of WFpkm between ATAC-seq experiments. Blue triangles, orangeXs, and green circles mark comparisons between independent experiments, between independent experimentsusing the same cell line, or true experimental replicates, respectively.

Additional file 8: Figure S8. The f1-scores, recall, and precision of the random forest model with ten-fold, stratified cross validation. Blue, orange, and green colorsdenote experimental relationship class.

Additional file 9: Table S1. Read counts of ATAC-seq experiments.

Additional file 10: Table S2. Fragment counts of ChIP-seq experiments from the ENCODE project.

Additional file 11: Table S3. Difference of mean correlation and association values between replicates and non-replicates.

Acknowledgements

We would like to thank members of the Genomics and Bioanalytics Group at Los Alamos National Laboratory for their helpful comments on this manuscript. Special thanks to Drs. Taehyung Kwon and Colin Kruse for their specific and helpful suggestions on analyses. We would also like to acknowledge and thank the ENCODE Consortium and the members of the Michael Snyder lab of Stanford, for generating the ATAC-seq data sets used within this study.

Author Contributions

The authors (with initials) CR, VV, VJ, NL, KYS, CRS, and SRS contributed to the overall experimental design of this manuscript. VV and CRS provided materials and wrote experimental methods. VV prepared experimental ATAC-seq data for cultured A549 cells. CR conducted analysis and produced visualizations. CR, CRS, and SRS wrote the paper. All authors edited and provided comments on the text of the manuscript.

Funding

The work conducted within this paper was funded by the Los Alamos National Laboratory Directed Research Grant (20210082DR) awarded to SRS and KYS. This work is approved for public release and distribution is unlimited under LA-UR-23-24317.

Availability of data and materials

All data and code associated with this manuscript is available upon request from the corresponding author, Cullen Roth (croth@lanl.gov). Raw sequence reads generated by this study on A549 samples are deposited on NCBI's Sequence Read Archive, under the bioproject PRJNA975595 with Accession Numbers SRR24717527–SRR24717534. Scripts, code, and software used in the statistical analysis and visualization are stored on GitHub: https://github.com/SLUR-m-Py/ATAC-seq_Simulation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 13 June 2023 Accepted: 30 October 2023
Published online: 22 November 2023

References

1. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823–37.
2. Barski A, Zhao K. Genomic location analysis by ChIP-Seq. J Cell Biochem. 2009;107(1):11–8.
3. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80.
4. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015;109(1):21–9.
5. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92.
6. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):1–9.
7. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–89.
8. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22(9):1813–31.
9. Oh D, Strattan JS, Hur JK, Bento J, Urban AE, Song G, et al. CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection. Sci Rep. 2020;10(1):7933.
10. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. 2020;21:1–16.
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.
12. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48(D1):D882–9.
13. Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. 2014;42(W1):W187–91.
14. Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc. 2022;17(6):1518–52.
15. Sahinyan K, Blackburn DM, Simon MM, Lazure F, Kwan T, Bourque G, et al. Application of ATAC-Seq for genome-wide analysis of the chromatin state at single myofiber resolution. Elife. 2022;11: e72792.
16. Zhao Y, Li MC, Konaté MM, Chen L, Das B, Karlovich C, et al. TPM, FPKM, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. J Transl Med. 2021;19(1):1–15.
17. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg. 2018;126(5):1763–8.
18. Milani P, Escalante-Chong R, Shelley BC, Patel-Murray NL, Xin X, Adam M, et al. Cell freezing protocol suitable for ATAC-Seq on motor neurons derived from human induced pluripotent stem cells. Sci Rep. 2016;6(1):1–10.
19. Shan X, Roberts C, Lan Y, Percec I. Age alters chromatin structure and expression of SUMO proteins under stress conditions in human adipose-derived stem cells. Sci Rep. 2018;8(1):11502.
20. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362(6413):eaav1898.
21. Halstead M, Kern C, Saelao P, Chanthavixay G, Wang Y, Delany M, et al. Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. Sci Rep. 2020;10(1):1–12.
22. Fang R, Preissl S, Li Y, Hou X, Lucero J, Wang X, et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun. 2021;12(1):1337.
23. Wong YY, Harbison JE, Hope CM, Gundsambuu B, Brown KA, Wong SW, et al. Parallel recovery of chromatin accessibility and gene expression dynamics from frozen human regulatory T cells. Sci Rep. 2023;13(1):5506.
24. Lynch M, Walsh B, et al. Genetics and analysis of quantitative traits, vol. 1. Sunderland: Sinauer; 1998.
25. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):1–19.
26. Yan KK, Yardımcı GG, Yan C, Noble WS, Gerstein M. HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps. Bioinformatics. 2017;33(14):2199–201.
27. Yang T, Zhang F, Yardımcı GG, Song F, Hardison RC, Noble WS, et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939–49.
28. Roth C, Sun S, Billmyre RB, Heitman J, Magwene PM. A high-resolution map of meiotic recombination in *Cryptococcus deneoformans* demonstrates decreased recombination in unisexual reproduction. Genetics. 2018;209(2):567–78.
29. Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. BMC Bioinform. 2018;19(1):1–10.

30. Yardımcı GG, Ozadam H, Sauria ME, Ursu O, Yan KK, Yang T, et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol. 2019;20(1):1–19.
31. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44(W1):W160–5.
32. Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat Commun. 2018;9(1):189.
33. Wolff J, Bhardwaj V, Nothjunge S, Richard G, Renschler G, Gilsbach R, et al. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. Nucleic Acids Res. 2018;46(W1):W11–6.
34. Wolff J, Rabbani L, Gilsbach R, Richard G, Manke T, Backofen R, et al. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. Nucleic Acids Res. 2020;48(W1):W177–84.
35. Nimon KF. Statistical assumptions of substantive analyses across the general linear model: a mini-review. Front Psychol. 2012;3:322.
36. Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 2020;18:2789–98.
37. Student. Probable error of a correlation coefficient. Biometrika. 1908;6(2-3):302–10.
38. Fisher R. Statistical methods for research workers Oliver and Boyd, London. Reprinted in Statistical Methods, Experimental Design and Scientific Inference; 1925.
39. Kowalski CJ. On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. J R Stat Soc Ser C (Appl Stat). 1972;21(1):1–12.
40. Kokoska S, Zwillinger D. CRC standard probability and statistics tables and formulae. CRC Press; 2000.
41. Kendall M. Rank correlation methods. 4th ed. High Wycombe, Bucks: Charles Griffin; 1970.
42. Noether GE. Elements of nonparametric statistics. Elements of nonparametric statistics; 1967.
43. Arndt S, Turvey C, Andreasen NC. Correlating and predicting psychiatric symptom ratings: Spearmans r versus Kendalls tau correlation. J Psychiatr Res. 1999;33(2):97–104.
44. Xu W, Hou Y, Hung Y, Zou Y. A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. Signal Process. 2013;93(1):261–76.
45. Cover TM. Elements of information theory. Wiley; 1999.
46. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
47. Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2012;2(6):493–507.
48. Iman RL, Conover W. A measure of top-down correlation. Technometrics. 1987;29(3):351–7.
49. Fitzgerald T, Jones A, Engelhardt BE. A Poisson reduced-rank regression model for association mapping in sequencing data. BMC Bioinform. 2022;23(1):1–22.
50. Kendall MG, Smith BB. The problem of m rankings. Ann Math Stat. 1939;10(3):275–87.
51. Burnham KP, Anderson DR, Burnham KP, Anderson DR. Practical use of the information-theoretic approach. Springer; 1998.
52. Varadan V, Miller DM III, Anastassiou D. Computational inference of the molecular logic for synaptic connectivity in *C. elegans*. Bioinformatics. 2006;22(14):e497–506.
53. Anastassiou D. Computational analysis of the synergy among multiple interacting genes. Mol Syst Biol. 2007;3(1):83.
54. Hu T, Chen Y, Kiralis JW, Collins RL, Wejse C, Sirugo G, et al. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. J Am Med Inform Assoc. 2013;20(4):630–6.
55. Budden DM, Crampin EJ. Information theoretic approaches for inference of biological networks from continuous-valued data. BMC Syst Biol. 2016;10(1):1–7.
56. Roth C, Murray D, Scott A, Fu C, Averette AF, Sun S, et al. Pleiotropy and epistasis within and between signaling pathways defines the genetic architecture of fungal virulence. PLoS Genet. 2021;17(1): e1009313.
57. Sun S, Roth C, Floyd Averette A, Magwene PM, Heitman J. Epistatic genetic interactions govern morphogenesis during sexual reproduction and infection in a global human fungal pathogen. Proc Natl Acad Sci. 2022;119(8): e2122293119.
58. Chen H, Lareau C, Andreani T, Vinyard ME, Garcia SP, Clement K, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. Genome Biol. 2019;20(1):1–25.
59. Xu Y, Das P, McCord RP. SMILE: mutual information learning for integration of single-cell omics data. Bioinformatics. 2022;38(2):476–86.
60. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.
61. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLoS ONE. 2016;11(11): e0167047.
62. Navidi Z, Zhang L, Wang B. simATAC: a single-cell ATAC-seq simulation framework. Genome Biol. 2021;22:1–16.
63. Giard DJ, Aaronson SA, Todaro GJ, Arnstein P, Kersey JH, Dosik H, et al. In vitro cultivation of human tumors: establishment of cell lines derived from a series of solid tumors. J Natl Cancer Inst. 1973;51(5):1417–23.
64. Foster KA, Oster CG, Mayer MM, Avery ML, Audus KL. Characterization of the A549 cell line as a type II pulmonary epithelial cell model for drug metabolism. Exp Cell Res. 1998;243(2):359–66.
65. Peng KJ, Wang JH, Su WT, Wang XC, Yang FT, Nie WH, et al. Characterization of two human lung adenocarcinoma cell lines by reciprocal chromosome painting. Dongwuxue Yanjiu. 2010;31(2):113–21.
66. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34(17):i884–90.
67. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44–53.
68. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

69. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30(17):2503–5.
70. Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011;9(4): e1001046.
71. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012;7(9):1728–40.
72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
73. Gaspar JM. Improved peak-calling with MACS2. BioRxiv. 2018;496521.
74. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann Appl Stat. 2011;5(3):1752–79.
75. Newell R, Pienaar R, Balderson B, Piper M, Essebier A, Bodén M. ChIP-R: assembling reproducible sets of ChIP-seq and ATAC-seq peaks from multiple replicates. Genomics. 2021;113(4):1855–66.
76. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
77. Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika. 1981;68(3):589–99.
78. John Lu Z. The elements of statistical learning: data mining, inference, and prediction; 2010.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.